

# Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk

Josep DOMINGO-FERRER<sup>1</sup>, Josep M. MATEO-SANZ<sup>1</sup> and Vicenç TORRA<sup>2</sup>  
<sup>1</sup>*Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics  
E-43006 Tarragona, Catalonia, Spain, E-mail {jdomingo,jmateo}@etse.urv.es*  
<sup>2</sup>*Institut d'Investigació en Intel·ligència Artificial, IIIA-CSIC, Campus de Bellaterra,  
E-08193 Bellaterra, Catalonia, Spain, E-mail vtorra@iia.csic.es*

**Abstract:** We present in this paper the first empirical comparison of SDC methods for microdata which encompasses both continuous and categorical microdata. Based on re-identification experiments, we try to optimize the tradeoff between information loss and disclosure risk. First, relevant SDC methods for continuous and categorical microdata are identified. Then generic information loss measures (not targeted to specific data uses) are defined, both in the continuous and the categorical case. Disclosure risk is assessed using empirical re-identification. Two approaches to empirical re-identification are used: Euclidean record linkage and probabilistic record linkage. The results of this comparison will be used to come up with better SDC for microdata in the recently started EU-funded project CASC.

**Keywords:** Statistical disclosure control, Microdata, Record linkage, Re-identification experiments, Information loss measures.

## 1. Introduction

This paper describes some of the results of the twin projects OTTILIE (Optimizing the Tradeoff between Information Loss and disclosure risk for microdata) sponsored by the U.S. Bureau of the Census (USBC). OTTILIE-R concentrates on comparing methods for continuous microdata and OTTILIE-D focuses on categorical microdata. The results obtained are useful for implementors and users of SDC for microdata, and especially for the recently started EU-funded CASC project.

The purpose of the experimentation work carried out under OTTILIE is to demonstrate a methodology for optimizing the tradeoff between information loss and disclosure risk. The approach used is as follows:

- *Literature analysis.* Literature on SDC for microdata has been analyzed to identify those methods which are relevant for protecting continuous and categorical microdata. Note that not all methods in the literature are useful for both kinds of microdata.
- *Test data.* Test data have been obtained from publicly available microdata files.
- *Information loss metrics.* Information loss actually depends on the data uses to be supported by the masked data. However, potential data uses are so diverse that it is hard even to identify them. The OTTILIE projects took a pragmatic approach to measure information loss, namely defining a battery of generic and simple information loss metrics that try to capture structural differences between the original and the masked data files.
- *Disclosure risk assessment.* This risk is empirically quantified. Two record linkage algorithms have been considered to establish the disclosure risk associated to a particular SDC method. In addition, a measure of interval disclosure has been taken into account.
- *Empirical work.* Experiments carried out are directed to obtaining t-uples of the form  $(method, parms, risk, loss)$ , where  $parms$  are the input parameters to  $method$ ,  $risk$  is the percent of re-identified records in the test data set and  $loss$  is the information loss. The obtained t-uples can be used to produce plots, tables and reports. Also, given  $risk$ , it is possible to find  $method$  and  $parms$  such that  $risk(loss, method(parms))$  is minimal (at least over the set of available t-uples). Finally, given  $loss$ , it should be possible to find  $method$  and  $parms$  such that  $risk(loss, method(parms))$  is minimal (at least over the set of available t-uples).

Section 2 reviews relevant SDC methods for the protection of continuous and categorical microdata. Section 3 lists information loss measures, both for the continuous and the categorical case. Section 4 describes record linkage approaches to assessing disclosure risk and defines also an interval disclosure measure. Section 5 reports on actual comparison results. Section 6 is a conclusion.

## 2. SDC methods included in the comparison

### 2.1 SDC methods for continuous microdata

Sampling methods consist of taking an unperturbed sample of the population microdata. Such methods may be suitable for categorical microdata, but their adequacy for continuous microdata is less clear in a general disclosure scenario. The reason is that a continuous variable is left unperturbed for all individuals in the sample. Thus, if variable  $V_i$  is present in an external administrative public file, unique matches are very likely, since for a

continuous variable (even a digitally represented one) it is highly unlikely that  $V_i(o_1) = V_i(o_2)$  if  $o_1 \neq o_2$ . Thus, we will consider only perturbative methods.

Perturbative methods distort records prior to release, which allows to release the whole population microdata set (not just a sample of it). Distortion is the best way to protect continuous microdata. Perturbative methods considered are a subset of those making sense for continuous microdata:

- *Additive noise* (Noise $p$  for short). Gaussian noise is added to the original data to get the masked data [Kim86]. If the standard deviation of the original variable is  $s$ , noise is generated using a  $N(0,ps)$ . Values of  $p$  considered in the experiments below are 0.01, 0.02, 0.04, 0.06, 0.08 up to 0.2 with 0.02 increments.
- *Data distortion by probability distribution* (Distr for short,[Liew85]). For each variable in the original data set, the best fitted distribution is found; then the fitted distribution is used to generate the masked data set. There are no parameters.
- *Resampling*. Take  $t$  independent samples  $X_1, \dots, X_t$  of the values of an original variable  $V_i$ . Sort all samples using the same ranking criterion. Build the masked variable  $V'_i$  by taking as first value the average of the first values of the samples, as second value the average of the second values and so on. Resampling has been tested for  $t=1$  (Resamp1) and  $t=3$  (Resamp3).
- *Microaggregation*. Records are clustered into small aggregates or groups of size at least  $k$  [Defa93,Domi02]. Rather than publishing a variable for a given individual, the average of the values of the variable over the group to which the individual belongs is published. Variants of microaggregation considered include: individual ranking (MicIR $k$ ); microaggregation on projected data using z-scores projection (MicZ $k$ ) and principal components projection (MicPC $k$ ); microaggregation on unprojected multivariate data considering two variables at a time (Mic2mul $k$ ), three variables at a time (Mic3mul $k$ ), four variables at a time (Mic4mul $k$ ) or all variables at a time (Micmul $k$ ). Values of  $k$  between 3 and 10 have been considered.
- *Lossy compression* (JPEG $q$ ). This method is new and proposed by these authors for continuous data. The idea is to regard a numerical microdata file as an image (with rows being records and columns being variables). Lossy compression, and more specifically the JPEG algorithm [JPEG], is then used on the image, and the compressed image is interpreted as a masked microdata file. Depending on the lossy compression algorithm used, appropriate mappings between variable ranges and color scales will be needed. The JPEG quality  $q$  has been taken as a parameter with values from 5% up to 100% with 5% increments.
- *Rank swapping* (Rank $p$ ). Although originally described only for ordinal variables, this method can be used for any numerical variable [Moor96]. First values of variable  $V_i$  are ranked in ascending order; then each ranked value of  $V_i$  is swapped with another ranked value randomly chosen within a restricted range (e.g. the rank of two swapped values cannot differ by more than  $p\%$  of the total number of records). Values of  $p$  from 1 to 20 have been considered in experimentation.

## 2.2 SDC methods for categorical microdata

For categorical data we have considered the methods and parameterizations given below. Each method depends on a single parameter and the set of variables to be masked. We describe the methods by Name-method, parameter and variables.

- *Top Coding*( $Tpv$  for short). This method is applied to ordinal categorical variables. In this case last  $p$  values of the variable are recoded into a new category. Values of  $p$  from 1 to 9 have been considered.
- *Bottom Coding*( $Bpv$  for short). This method is also applied to ordinal categorical variables. In this case first  $p$  values of the variable are recoded into a new category. Values of  $p$  from 1 to 9 have been considered.
- *Global recoding* ( $Gpv$  for short). In global recoding, some of the categories of the variable are recoded into new ones. In our case, we have considered the following parameterization: the  $p$  categories with a lower frequency are recoded into a single variable. Values of  $p$  from 1 to 9 have been considered.
- PRAM ( $Ppv$  for short). The scores of some categorical variables for certain records in the original file are changed to a different score according to a prescribed probability mechanism (a Markov matrix). In our case, we have selected the approach described in [KWG98] to define the PRAM matrix. This is as follows: Let  $T_V=(T_V(1), \dots, T_V(K))'$  be the K-vector of frequencies in the original file of the  $K$  categories of the categorical variable  $V$  (assume without loss of generality that  $T_V(k) \geq T_V(K) > 0$  for  $k < K$ ) and let  $\theta$  be such that  $0 < \theta < 1$ , then the PRAM matrix for variable  $V$  is defined as follows:

$$P_{kl} = \begin{cases} 1 - \theta T_V(K) / T_V(k) & \text{if } l = k \\ \theta T_V(K) / ((K - 1)T_V(k)) & \text{if } l \neq k \end{cases}$$

For each variable we have built 9 matrices generated with an integer  $p$  from 1 to 9. This has been obtained using the approach described above defining  $\theta$  as  $p$  divided by 10.

## 3. Information loss measures

To evaluate the information loss caused by an SDC method on a microdata set, we want to assess how different the masked data set is from the original data set. We will say there is little information loss if the structure of the masked data set is very similar to the structure of the original data set. In fact, the motivation for preserving the structure of the data set is to ensure that the masked data set will be analytically valid and interesting. We can actually try several complementary ways to assess the preservation of the structure of the original data set:

Compare the data in the original and the masked data sets. The more similar the SDC method to the identity function, the less impact (but the higher the disclosure risk!).

Compare some statistics computed on the original and the masked data sets.

### 3.1 Information loss measures for continuous microdata

Let  $X$  and  $X'$  be the original and the masked data set. Let  $V$  and  $V'$  be the covariance matrices of  $X$  and  $X'$ , respectively; similarly, let  $R$  and  $R'$  be the correlation matrices. Table 1 summarizes the measures proposed. In this table,  $p$  is the number of variables,  $n$  the number of records, and components of matrices are represented by the corresponding lowercase letters (e.g.  $x_{ij}$  is a component of matrix  $X$ ). Regarding  $X - X'$  measures, it also makes sense to compute those on the averages of variables rather than on all data (see the  $\bar{X} - \bar{X}'$  row in Table 1). Similarly, for  $V - V'$  measures, it is also sensible to compare only the variances of the variables, *i.e.* to compare the diagonals of the covariance matrices rather than the whole matrices (see the  $S - S'$  row in Table 1).

**Table 1.** Information loss measures

	Mean square error	Mean abs. Error	Mean variation
$X - X'$	$\frac{\sum_{j=1}^p \sum_{i=1}^n (x_{ij} - x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^p \sum_{i=1}^n  x_{ij} - x'_{ij} }{np}$	$\frac{\sum_{j=1}^p \sum_{i=1}^n \frac{ x_{ij} - x'_{ij} }{ x_{ij} }}{np}$
$\bar{X} - \bar{X}'$	$\frac{\sum_{j=1}^p (\bar{x}_j - \bar{x}'_j)^2}{p}$	$\frac{\sum_{j=1}^p  \bar{x}_j - \bar{x}'_j }{p}$	$\frac{\sum_{j=1}^p \frac{ \bar{x}_j - \bar{x}'_j }{ \bar{x}_j }}{p}$
$V - V'$	$\frac{\sum_{j=1}^p \sum_{1 \leq i \leq j} (v_{ij} - v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{1 \leq i \leq j}  v_{ij} - v'_{ij} }{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{1 \leq i \leq j} \frac{ v_{ij} - v'_{ij} }{ v_{ij} }}{\frac{p(p+1)}{2}}$
$S - S'$	$\frac{\sum_{j=1}^p (v_{jj} - v'_{jj})^2}{p}$	$\frac{\sum_{j=1}^p  v_{jj} - v'_{jj} }{p}$	$\frac{\sum_{j=1}^p \frac{ v_{jj} - v'_{jj} }{ v_{jj} }}{p}$

R-R'	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j} (r_{ij} - r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j}  r_{ij} - r'_{ij} }{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{1 \leq i < j} \frac{ r_{ij} - r'_{ij} }{ r_{ij} }}{\frac{p(p-1)}{2}}$
------	--	--	---

### 3.2 Information loss measures for categorical microdata

For categorical data three kinds of information loss measures have been considered: direct comparison of categorical values, comparison of contingency tables and entropy-based measures.

**Direct comparison of categorical values:** A distance was defined over the range of categorical variables. When the range of a variable is an ordinal scale, the distance between category  $a$  and  $b$  is proportional to the number of categories between  $a$  and  $b$ . When the range of a variable is not ordinal, the distance is one if the values are different and zero if they are not. We denote this information loss measure by Dist.

**Comparison of contingency tables:** For a given subset of variables, contingency tables are computed for a file before and after applying the masking process. The number of differences between both contingency tables is denoted by CTBIL (Contingency Table Based Information Loss measure). As the number of cells in a contingency table depends on the number of categories in the variable, we have also considered a normalized expression of CTBIL. This corresponds to the former information loss measure but dividing the expression by the number of cells in all considered tables. We express this information loss measure by A CTBIL (Average CTBIL).

**Entropy-based measures:** In [KWG98] the use of Shannon's entropy to measure information loss is discussed. The idea is that this information-theoretic measure can be used in SDC if the masking process is modeled as the noise that would be added to the original data set in the event of it being transmitted over a noisy channel. As this measure only depends on the masked data set and it does not account for its relation with the original data, a new information loss measure was defined.

Let  $V$  be a variable in the original data set and  $V'$  be the corresponding variable in the PRAM-masked data set (we take PRAM because it is a very general method encompassing the rest of masking methods considered for categorical data). Then, the entropy-based information loss measure  $EBIL$  is defined as:

$$EBIL(P_{V,V'}, G) = \sum_{r \in G} H(V|V'=j_r)$$

where  $j_r$  is the value taken by  $V'$  in record  $r$ , and

$$H(V|V' = j) = -\sum_{i=1}^n p(V = i|V' = j) \log p(V = i|V' = j)$$

$P_{V,V'} = \{p(V'=j|V=i)\}$  being the PRAM Markov matrix.

The new information loss measure taking original data into account is:

$$IL(P_{V,V'}, F, G) = \sum_{r \in G} PRIL(P_{V,V'}, i_r, j_r)$$

where  $i_r$  is the value taken by  $V$  in record  $r$  of  $F$  and  $j_r$  is, as before, the value taken by  $V'$  in record  $r$  of  $G$  and

$$PRIL(P_{V,V'}, i, j) = -\log P(V=i|V'=j)$$

#### 4. Disclosure risk measures

The assessment of the quality of an SDC method cannot be limited to information loss; disclosure risk is another magnitude that should be measured. The method that optimizes the tradeoff between both magnitudes subject to some user requirements turns out to be the best option.

Literature on disclosure risk is basically related to sampling methods, in which a sample of the original data set is published. Disclosure risk here is measured as the probability that a sample unique is a population unique [Skin94]. If the size of the sample is similar to the size of the whole population, such a probability can be dangerously high; in that case, an intruder who locates a unique value in the released sample could be almost sure that there is a single individual in the population with that value. This could lead to identification of that individual.

The uniqueness property as stated above is no longer relevant for perturbative methods, since in this case the whole microdata set is published, but with some distortion. There is not much literature on disclosure risk that can be used for a broad class of perturbative methods; disclosure risk measures tend to be method-specific (measures described in [Adam89] are still up-to-date). Empirical methods, like record linkage techniques, provide a more unified approach to disclosure risk assessment for perturbative methods. We briefly describe below two approaches to record linkage and one measure of interval disclosure.

##### 4.1 Distance-based record linkage

This approach to record linkage is described in [Pagl98] for the specific case of microaggregation masking and using the Euclidean distance. However, it can be generalized for any perturbative method provided that a distance between the original and the masked value can be defined. As in any record linkage context, it is assumed that an intruder has an external data set containing as key variables some of the same variables present in the released masked data set. The intruder is assumed to try to link the masked data set with the external data set.

Linkage then proceeds by computing the distances between records in the original and the masked data sets. The distances used are standardized to avoid scaling problems. For each record in the masked data set, the distance to every record in the original data set is computed. Then the "nearest" and "second nearest" records in the original data set are considered. A record in the masked data set is labelled as "linked" when the nearest record in the original data set has the same record number is the corresponding original record). A record in the masked data set is labelled as "linked to 2nd nearest" when the second nearest record in the original data set has the same record number. In all other cases, a record in the masked data set is labelled as "not linked". The percent of "linked" and "linked to 2nd nearest" is a measure of disclosure risk.

#### **4.2 Probabilistic record linkage**

In [Jaro89], a probabilistic record linkage method was described and illustrated on the 1985 Census of Tampa, Florida. The matching algorithm uses the linear sum assignment model to "pair" records in the two files to be matched (the original file and the masked file in our case). The percent of correctly paired records is a measure of disclosure risk.

Although less simple than the Euclidean method described in the previous section, this approach is attractive because it only requires the user to provide two probabilities as input: one is an upper bound of the probability of a false match, and the other an upper bound of the probability of false non-match. The Euclidean method above requires rescaling variables as well as an assumption on the weight of variables when computing a distance: for instance, in the proposal of [Pagl98], all variables have the same weight.

The U.S. Census Bureau implementation of probabilistic record linkage provided by W. Winkler [USBC,Wink98] has been used (with some additions) in the experimentation.

#### **4.3 Interval disclosure**

For a record in the masked data set, take a rank interval centered on the values of that record as follows: each variable is independently ranked and a rank interval is defined around the value the variable takes on each record; the ranks of values within the interval for a variable around record  $r$  should differ less than  $p\%$  of the total number of records and the rank in the center of the interval should correspond to the value of the variable in record  $r$ . Then the measure is the proportion of original values which fall into the interval centered around their corresponding masked value. A 100% proportion means that an intruder is completely sure that the original value lies in the interval around the masked value (interval disclosure). Values of  $p$  ranging between 1% and 10% have been considered for experimentation.



## 5. Comparison results

### 5.1 Comparison for continuous microdata

A microdata set was constructed using the Data Extraction System (DES) of the U.S. Census Bureau (<http://www.census.gov/DES>). 13 continuous variables were chosen and 1080 records were selected so that there were not many repeated values for any of the variables (in principle, one would not expect repeated values for a continuous variable, but there were repetitions in the data set). Table 3 contains a ranking of methods described in Section 2 (the parameter values described in that section were tried for each method). The Information Loss column (IL) is computed by averaging the mean variations of  $X-X'$ ,  $\bar{X} - \bar{X}'$ ,  $V-V'$ ,  $S-S'$  and the mean absolute error of  $R-R'$ ; the resulting average has been multiplied by 100. The Distance Linkage Disclosure risk column (DLD) contains the average percent of linked records using distance-based record linkage; the average is computed over the number of key variables that the intruder is assumed to know (we have considered knowledge of 1 up to 7 variables). Similarly, the Probabilistic Linkage Disclosure risk column (PLD) is the average percent of correctly paired records using probabilistic linkage. The Interval Disclosure (ID) column contains the average percent of original values falling in the intervals around their corresponding masked values (averages have been computed over all parameter values, i.e. 1% to 10% with 1% increments). Finally, the column Score has been used to rank Table 3 and has been computed as

$$\text{Score} = 0.5(\text{IL}) + 0.125(\text{DLD}) + 0.125(\text{PLD}) + 0.25(\text{ID}).$$

The rationale of the above weighting is to give equal weight to information loss (0.5) and to disclosure risk. The 0.5 weight of disclosure risk is equally divided among ID (0.25) and record linkage. The 0.25 weight of record linkage is equally divided among both approaches to record linkage. The correlation between DLD and PLD is actually 0.962, so both approaches are very similar. The (IL, DLD), (IL, PLD) and (IL, ID) correlations are  $-0.605$ ,  $-0.551$  and  $-0.807$ ; thus, the lower the information loss, the higher the disclosure risk, as one would expect. The IL Rank, DLD Rank, PLD Rank and ID Rank columns contain the ranking of each method with respect to IL, DLD, PLD and ID; the lower the rank, the better a method performs (i.e. lower information loss and disclosure risk).

Note that publishing the original data without masking yields score 50 (IL=0 and DLD=PLD=ID=100). Therefore, methods scoring above 50 in Table 3 are of little use.

### 5.2 Comparison for categorical microdata

For comparing the masking methods for categorical microdata, we have used data from the *American Housing Survey 1993* (also obtained from the U.S. Census Bureau using the Data Extraction System - DES). We selected the variables BUILT, DEGREE, GRADE1,

METRO, SCH, SHP, TRAN1, WHYMOVE, WHYTOH, WHYTON. From the corresponding data file, we have selected the first 1000 records.

Five subsets of variables were defined from the set of selected variables, and for each of them the same analysis has been performed in the testing process. Three groups were defined grouping the categories with similar number of categories. "m", "z" and "p" correspond to categories with a few, medium and large number of labels ("m" stands for minus, "z" for zero and "p" for positive). Additionally, "g" corresponds to the union of variables in "z" and "p"; and "o" corresponds to the subset of variables that were defined as ordered. The variables used and the groups of variables are given in Table 2. This Table also includes the number of categories for each variable.

Variables	g	p	m	z	o	N. of Cat.
BUILT	X	X			X	25
DEGREE			X		X	8
GRADE1	X	X			X	21
METRO			X			9
SCH			X			6
SHP			X			6
TRAN1	X			X		12
WHYMOVE	X	X				18
WHYTOH	X			X		13
WHYTON	X			X		13

**Table 2.** Variables used in the masking process and in the reidentification process.

Table 4 contains the rank of masking methods and the results obtained for each experiment. Each row corresponds to a different experiment (method, parameter and variables) and for each one the following information is included:

- PLD: number of reidentified records (over 1000) using Probabilistic Record Linkage
- Dist, CTBIL A CTBIL, EBIL and IL: information loss measures. EBIL and IL have been computed using two different data files to estimate probabilities: the same file we are masking (File to M) and a reference file (Ref F)
- PLD rank: rank according to Probabilistic Record Linkage (normalized, maximum 100)
- Ave IL R: rank resulting from the average ranks for all information loss measures. This has been computed to give the same weight to the three information loss measures defined in Section 3.2: distance, contingency table and entropy based measures. Each group has the same importance and within a group all the measures also have the same importance. This results into the following expression:  

$$\frac{(\text{Dist} + (\text{CTBIL} + \text{ACTBIL})/2 + (\text{EBIL Ref F} + \text{IL Ref F} + \text{EBIL File to M} + \text{IL File to M})/4)}{3}$$
 This has been normalized into the [0,100] interval.
- Score: the average of the two previous ranks.
- Ave. Score: the average score for all the experiments that use the same parameterizations (except for the set of variables).

Results presented here are based on ranks instead of on the values of the information loss measures because the range of the latter are different and difficult to compare among each other. The correlation between information loss measures and the probabilistic record linkage are as follows:

DIST	CTBIL	ACTBIL	Ent Ref F	IL Ref F	Ent File to M	IL File to M
-0.4898	-0.4156	-0.5520	-0.345	-0.288	-0.368	-0.408

Here again, as expected, the lower the information loss, the higher the disclosure risk.

## 6. Conclusions

There is a rich array of methods for microdata disclosure limitation. A set of proposals for continuous microdata have been identified and described in this paper. Measures for assessing information loss have also been described.

Regarding methods for masking continuous microdata (Table 3), Distr, MicZk, MicIRk, MicPCPk and Resampling score around or above 50 for all tried  $k$ , so their use is not recommended. Multivariate microaggregation on unprojected data, proposed in [Domi02], is the only form of microaggregation scoring well. Taking three variables at a time seems the best strategy but, even in this case, microaggregation is second to rank swapping. Rank swapping outperforms the best microaggregation and is the best performer (for  $p$  around 15%). However, being a stochastic method, it is not reproducible and this may lead to disclosure in on-line databases allowing repeated queries. Therefore, microaggregation on unprojected data can still be a good option in such cases. Lossy compression (JPEG) is not excellent but is promising; being a new proposal, there is room for improvement.

Regarding methods for categorical microdata, parameterizations of top coding are the best-rated methods while the PRAM methods (with the parameters described above) are poorly rated. In general, for the former methods the re-identification rate is low while the information loss is moderate and for the latter methods, the re-identification is high while the information loss is also high. However, the results show that the information loss and the number of re-identifications are highly dependent on the set of variables and the number of categories in each variable. In this sense, the selected parameterization for PRAM seems particularly inconvenient for variables with a large number of categories.

## Acknowledgments

This work was partly funded by the U.S. Bureau of the Census under contracts no OBLIG-2000-29158-0-0 and OBLIG-2000-29144-0-0 and by the European Commission under project "CASC" IST-2000-25069. Thanks go to Francesc Seb , Narc s Maci  and  ngel Torres for their help in automating the probabilistic record linkage software and running the experiments.

**References**

- [1] Adam, N. R., Wortmann, J. C., (1989), Security-control methods for statistical databases: a comparative study, *ACM Computing Surveys*, vol. 21(4):515-556.
- [2] Defays, D., Nanopoulos, P., (1993), Panels of enterprises and confidentiality: the small aggregates method, in *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys*, Ottawa: Statistics Canada, 195-204.
- [3] Domingo-Ferrer, J., Mateo-Sanz, J.M., (2002), Practical Data-Oriented Microaggregation for Statistical Disclosure Control, *IEEE Transactions on Knowledge and Data Engineering*, (to appear, March 2002).
- [4] Jaro, M. A., (1989), Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, vol. 84:414-420.
- [5] Joint Photographic Experts Group, Standard IS 10918-1 (ITU-T T.81) <http://www.jpeg.org>.
- [6] Kim, J. J., (1986), A method for limiting disclosure in microdata based on random noise and transformation, in *Proc. of the ASA Sect. on Survey Res. Meth.*, pp. 303-308.
- [7] Kooiman, P., Willenborg, L., Gouweleew, J., (1998), PRAM: a method for disclosure limitation of microdata, Research Report, Statistics Netherlands.
- [8] Liew, C. K., Choi, U. J., Liew, C. J., (1985), A data distortion by probability distribution, *ACM Transactions on Database Systems*, vol. 10: 395-411.
- [9] Moore, R., (1996), Controlled data swapping techniques for masking public use microdata sets, U. S. Bureau of the Census (unpublished manuscript).
- [10] Pagliuca, D., Seri, G., (1998), Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, Deliverable MI-3/D2.
- [11] Skinner, C., Marsh, C., Openshaw, S., Wymer, C., (1994), Disclosure Control for Census Microdata, *Journal of Official Statistics*, vol. 10:31-51.
- [12] U. S. Bureau of the Census, (2000), Record Linkage Software: User Documentation. Available from U. S. Bureau of the Census.
- [13] Winkler, W., (1998), Re-identification methods for evaluating the confidentiality of analytically valid microdata, in *Statistical Data Protection*, Luxembourg: Office for Official Publications of the European Communities, 1999. Journal version in *Research in Official Statistics*, vol. 1(2): 50-69, 1998.

**Table 3.** Comparison results for continuous microdata

Method	IL	DLD	PLD	ID	Score	IL Rank	DLD Rank	PLD Rank	ID Rank
Rank15	19.01	1.19	0.15	35.05	18.44	53	6	7	21
Rank19	22.95	0.93	0.08	28.04	18.61	59	2	2	2
Rank16	20.91	1.39	0.11	32.18	18.69	56	8	5	16
Rank13	16.77	2.17	0.12	40.35	18.76	48	12	6	28
Rank14	19.72	1.92	0.07	37.00	19.36	55	10	1	25
Rank11	14.32	2.43	0.25	47.81	19.45	44	13	14	39
Rank12	16.37	2.50	0.25	43.73	19.46	47	14	11	35
Rank20	25.81	0.69	0.09	26.83	19.71	64	1	3	1
Rank18	25.74	0.95	0.09	29.25	20.31	63	4	4	6
Rank10	13.37	3.90	0.38	53.17	20.51	41	24	17	45
Rank17	25.12	1.52	0.20	30.95	20.51	61	9	9	10
Rank09	11.66	5.01	0.52	57.58	20.91	38	37	29	49
Rank08	11.60	6.07	0.85	63.37	22.51	37	39	39	56
Rank07	9.25	7.51	1.08	68.71	22.87	30	41	43	63
Rank06	7.87	9.02	2.79	73.80	23.86	26	43	56	71
Mic3mul07	11.06	19.34	4.70	72.34	26.62	36	68	65	69

Rank05	6.78	16.80	13.60	78.89	26.91	22	58	70	77
Mic3mul09	13.46	19.22	3.44	69.91	27.04	42	67	60	65
Mic3mul10	14.84	17.99	3.44	68.61	27.25	46	64	59	62
Mic4mul04	12.14	19.76	6.67	71.85	27.33	39	69	68	68
Mic4mul05	14.50	17.43	5.45	69.09	27.39	45	61	66	64
Mic3mul08	13.51	20.81	4.15	70.68	27.54	43	71	63	66
Mic4mul08	18.89	17.78	3.35	62.84	27.80	52	62	58	55
Mic3mul06	10.24	20.41	13.90	74.00	27.91	33	70	71	72
Mic4mul07	19.36	17.10	2.08	64.41	28.18	54	60	53	58
Mic4mul06	17.91	17.82	3.98	66.41	28.28	50	63	62	60
Mic4mul09	21.35	15.93	2.00	61.66	28.33	58	57	52	54
Mic4mul10	22.98	16.85	2.37	60.56	29.03	60	59	55	51
Mic3mul05	9.73	23.78	18.29	76.59	29.27	31	76	73	74
Mic3mul04	7.45	23.49	22.75	79.14	29.29	24	75	75	79
Mic4mul03	10.69	22.88	16.69	76.89	29.51	35	74	72	75
Rank04	5.90	22.77	22.78	84.12	29.67	20	73	76	86
Micmul03	27.67	14.26	1.88	57.23	30.16	65	54	50	47
Micmul04	31.74	13.72	1.38	52.44	30.86	67	53	48	44
Mic3mul03	6.29	29.70	29.06	82.95	31.23	21	79	80	85
Micmul05	35.12	11.73	1.14	48.43	31.27	70	46	44	41
Micmul07	37.68	13.20	1.20	43.46	31.50	72	52	45	34
Micmul06	38.77	13.00	1.22	45.76	32.60	73	50	46	37
Micmul08	41.53	13.12	0.99	42.66	33.19	75	51	42	32
Rank03	5.07	31.73	36.92	89.53	33.50	18	80	83	93
Mic2mul10	10.68	49.38	27.29	77.43	34.28	34	86	78	76
Micmul10	44.69	14.66	0.50	40.41	34.34	76	55	27	29
Noise0.16	32.56	15.65	4.66	64.39	34.91	68	56	64	57
Micmul09	45.98	12.82	0.85	40.99	34.95	79	49	40	30
Mic2mul09	9.93	51.03	33.04	78.94	35.21	32	87	81	78
Mic2mul08	8.55	54.31	33.70	79.77	35.22	27	88	82	80
Mic2mul07	7.53	54.72	37.41	81.40	35.63	25	89	84	83
Noise0.12	25.24	22.21	22.39	71.58	36.09	62	72	74	67
Noise0.1	21.14	27.70	29.03	75.20	36.46	57	78	79	73
Mic2mul06	7.03	56.38	42.00	82.89	36.54	23	90	86	84
JPEG080	33.97	19.13	6.93	66.35	36.83	69	65	69	59
Noise0.14	35.13	19.21	6.24	67.62	37.65	71	66	67	61
Noise0.18	41.12	11.96	3.52	60.95	37.73	74	47	61	52
Noise0.08	17.43	36.06	39.76	79.84	38.15	49	82	85	81
Rank02	2.90	47.26	57.47	94.56	38.18	11	85	90	96
JPEG070	44.92	9.66	2.34	57.28	38.28	77	44	54	48
Noise0.2	45.97	10.01	0.97	57.63	38.77	78	45	41	50
Mic2mul05	5.88	58.97	56.84	85.40	38.77	19	92	89	88
JPEG085	29.47	23.85	24.48	72.80	38.98	66	77	77	70

Mic2mul04	4.90	61.53	60.69	87.26	39.54	17	94	91	89
JPEG090	18.17	35.37	46.98	80.87	39.60	51	81	87	82
Noise0.06	13.03	45.54	56.22	84.16	40.28	40	84	88	87
Mic2mul03	3.28	66.97	64.79	90.51	40.74	15	95	92	94
Noise0.04	8.93	58.51	65.28	88.95	42.18	28	91	94	90
JPEG075	50.45	12.67	2.90	61.27	42.49	80	48	57	53
JPEG095	9.06	60.11	66.56	89.23	42.67	29	93	96	92
Resamp3	3.15	67.90	67.63	96.81	42.72	14	96	97	97
Rank01	2.34	69.19	66.35	99.54	43.00	9	97	95	106
JPEG065	57.77	7.02	1.90	53.87	43.47	81	40	51	46
Noise0.02	4.24	77.34	71.32	94.42	44.31	16	99	98	95
Resamp1	3.11	75.42	71.85	98.36	44.56	13	98	99	99
MicPCP03	69.62	3.16	0.77	38.41	44.90	84	17	38	26
JPEG055	63.70	5.57	1.26	49.70	45.13	83	38	47	42
Noise0.01	2.57	85.19	74.13	97.03	45.46	10	100	103	98
JPEG100	3.06	87.14	73.03	99.14	46.34	12	101	100	101
MicIR10	1.19	97.37	74.07	99.12	46.81	8	102	102	100
MicIR08	1.03	97.84	74.07	99.29	46.83	6	108	101	103
MicIR09	1.14	97.96	74.40	99.24	46.93	7	109	104	102
MicIR06	0.87	97.66	75.28	99.51	46.93	5	106	105	105
MicIR05	0.69	97.58	75.99	99.58	46.94	3	104	106	107
MicIR03	0.45	97.39	78.96	99.79	47.22	1	103	107	109
MicIR04	0.64	97.63	79.78	99.67	47.41	2	105	108	108
MicIR07	0.81	97.79	88.06	99.42	48.49	4	107	109	104
MicPCP04	78.84	3.43	0.62	36.00	48.92	87	19	32	23
JPEG050	73.20	4.26	0.67	47.96	49.21	86	31	36	40
JPEG060	71.24	7.66	1.52	51.71	49.69	85	42	49	43
MicPCP05	82.55	3.94	0.69	34.10	50.38	88	25	37	20
MicPCP07	89.28	4.02	0.62	32.56	53.36	91	27	33	17
MicPCP09	90.78	4.54	0.25	31.40	53.84	94	34	12	13
MicPCP06	90.26	3.37	0.50	33.42	53.97	93	18	26	19
MicZ03	90.25	3.16	0.61	35.71	54.52	92	16	31	22
JPEG035	88.80	3.65	0.44	43.20	55.71	90	20	23	33
JPEG045	87.55	4.15	0.67	46.78	56.07	89	30	35	38
MicZ04	94.94	3.70	0.53	33.04	56.26	96	21	30	18
MicPCP08	96.93	3.97	0.34	32.04	57.02	97	26	16	14
MicPCP10	97.82	4.13	0.46	31.19	57.28	98	29	24	11
JPEG040	90.99	3.72	0.66	44.98	57.29	95	22	34	36
MicZ07	102.87	4.27	0.38	30.53	59.65	99	32	20	9
MicZ06	103.92	3.88	0.41	30.43	60.10	100	23	21	8
MicZ05	104.06	4.03	0.42	31.30	60.41	101	28	22	12
MicZ08	107.92	4.55	0.52	29.60	61.99	102	35	28	7
MicZ10	109.79	4.83	0.38	28.20	62.59	103	36	18	3

MicZ09	110.91	4.35	0.38	28.36	63.14	105	33	19	4
Distr	58.62	43.05	64.88	88.98	65.04	82	83	93	91
JPEG030	110.48	3.02	0.48	41.79	66.12	104	15	25	31
JPEG025	155.15	2.13	0.25	38.76	87.56	106	11	13	27
JPEG020	164.91	1.36	0.29	36.11	91.69	107	7	15	24
JPEG015	202.66	1.10	0.15	32.06	109.50	108	5	8	15
JPEG010	269.38	0.93	0.22	28.44	141.94	109	3	10	5

**Table 4.** Comparison results for categorical microdata

Method	PLD	Dist	CTBIL	A CTBIL	EBIL Ref. F	IL Ref. F	EBIL File to M	IL File to M	PLD Rank	Ave IL R.	Score	Ave. Score
T5g	235	2716	31830	11.1	2892.1	3030.3	2952.6	2952.6	5.6	72.82	39.19	41.60
T5m	428	2372	14788	27.0	2453.4	2564.1	2506.9	2506.9	23.9	70.19	47.04	41.60
T5o	347	1093	6148	5.2	1661.0	1727.4	1687.5	1687.5	15.0	56.76	35.88	41.60
T5p	853	319	1820	1.0	486.8	481.4	474.9	474.9	58.3	32.36	45.35	41.60
T5z	769	344	1874	2.7	438.7	466.2	445.7	445.7	45.0	36.11	40.56	41.60
T6g	200	3152	35786	12.5	3764.1	3900.7	3819.5	3819.5	5.0	80.23	42.62	42.19
T6m	457	2789	16400	30.0	3256.0	3357.3	3299.8	3299.8	26.7	74.81	50.74	42.19
T6o	287	1242	6886	5.8	2106.7	2171.9	2130.5	2130.5	10.0	59.77	34.88	42.19
T6p	789	391	2220	1.3	671.2	663.6	655.5	655.5	48.9	37.18	43.03	42.19
T6z	751	363	1966	2.9	508.1	543.4	519.7	519.7	41.7	37.64	39.65	42.19
T3g	288	1081	14154	4.9	942.1	963.8	946.4	946.4	10.6	55.65	33.10	42.90
T3m	748	819	6150	11.2	691.2	717.4	706.7	706.7	41.1	51.62	46.37	42.90
T3o	758	559	3258	2.8	576.4	588.4	579.1	579.1	42.8	41.90	42.34	42.90
T3p	943	177	1028	0.6	179.4	178.9	173.3	173.3	78.9	23.38	51.13	42.90
T3z	825	262	1442	2.1	250.9	246.5	239.7	239.7	52.8	30.37	41.57	42.90
T4g	254	2475	29456	10.3	2289.3	2357.7	2311.4	2311.4	7.2	69.07	38.15	43.15
T4m	492	2144	13756	25.1	1896.4	1943.9	1917.6	1917.6	28.3	67.27	47.80	43.15
T4o	529	854	4896	4.1	1117.7	1135.1	1124.9	1124.9	30.0	50.65	40.32	43.15
T4p	908	233	1348	0.8	304.9	303.1	297.1	297.1	71.1	27.31	49.21	43.15
T4z	775	331	1812	2.6	392.9	413.8	393.8	393.8	46.1	34.40	40.25	43.15
G1g	423	55	770	0.3	0.0	0.0	0.0	0.0	23.3	12.41	17.87	43.25
G1m	1000	51	408	0.9	0.0	0.0	0.0	0.0	98.3	13.80	56.06	43.25
G1o	998	35	210	0.2	0.0	0.0	0.0	0.0	95.6	6.48	51.02	43.25
G1p	998	4	24	0.0	0.0	0.0	0.0	0.0	96.1	0.65	48.38	43.25
G1z	962	4	24	0.0	0.0	0.0	0.0	0.0	84.4	1.39	42.92	43.25
T7g	265	3453	38344	13.4	4389.9	4564.0	4475.6	4475.6	7.8	82.64	45.21	44.15
T7m	687	3067	17418	31.8	3801.9	3943.7	3879.9	3879.9	35.0	80.09	57.55	44.15
T7o	307	1314	7172	6.1	2281.9	2345.2	2303.7	2303.7	12.8	61.76	37.27	44.15

T7p	719	483	2736	1.6	903.2	893.8	885.5	885.5	38.3	41.71	40.02	44.15
T7z	751	386	2072	3.0	588.0	620.3	595.7	595.7	42.2	39.17	40.69	44.15
T9g	429	4163	43976	15.4	4921.4	5230.9	5133.5	5133.5	24.4	88.15	56.30	44.18
T9m	187	3695	19294	35.3	4097.7	4368.1	4305.5	4305.6	4.4	83.56	44.00	44.18
T9o	297	1604	8162	6.9	2925.9	2917.3	2834.9	2834.9	12.2	66.99	39.61	44.18
T9p	534	812	4392	2.5	1665.6	1585.9	1536.0	1536.0	30.6	49.35	39.95	44.18
T9z	725	468	2442	3.5	823.7	862.7	828.0	828.0	38.9	43.19	41.04	44.18
T8g	510	4167	43940	15.3	4957.2	5244.2	5140.2	5140.2	28.9	88.33	58.61	44.75
T8m	44	3774	19386	35.4	4337.2	4583.7	4511.2	4511.2	0.6	84.44	42.50	44.75
T8o	329	1494	7772	6.6	2635.5	2622.5	2542.5	2542.6	14.4	64.81	39.63	44.75
T8p	643	679	3726	2.1	1308.4	1222.4	1175.5	1175.5	33.3	47.08	40.21	44.75
T8z	774	393	2108	3.1	620.0	660.5	629.0	629.0	45.6	40.00	42.78	44.75
G4g	388	1269	16606	6.9	1246.1	1272.4	1241.8	1241.8	16.7	60.88	38.77	45.35
G4m	729	1215	8690	19.8	1183.5	1203.2	1175.6	1175.6	39.4	61.48	50.46	45.35
G4o	810	424	2508	2.4	560.1	597.2	517.5	517.5	51.1	39.21	45.16	45.35
G4p	983	71	420	0.3	81.9	144.9	76.1	76.1	90.0	14.63	52.31	45.35
G4z	893	54	316	0.5	62.6	69.3	66.1	66.1	66.1	13.94	40.02	45.35
T2g	324	655	8802	3.1	406.5	400.4	388.9	388.9	13.9	44.35	29.12	45.36
T2m	890	462	3598	6.6	301.4	299.5	293.3	293.3	65.6	39.72	52.64	45.36
T2o	903	330	1960	1.7	218.0	210.5	205.9	205.9	68.3	30.79	49.56	45.36
T2p	965	132	780	0.4	83.2	79.4	76.1	76.1	85.0	19.49	52.25	45.36
T2z	873	193	1058	1.5	105.0	101.0	95.6	95.6	61.1	25.37	43.24	45.36
T1g	397	222	3058	1.1	0.0	0.0	0.0	0.0	19.4	25.00	22.22	46.88
T1m	1000	172	1362	2.5	0.0	0.0	0.0	0.0	98.9	23.70	61.30	46.88
T1o	1000	109	654	0.6	0.0	0.0	0.0	0.0	99.4	15.56	57.50	46.88
T1p	996	40	240	0.1	0.0	0.0	0.0	0.0	93.9	6.30	50.09	46.88
T1z	933	50	272	0.4	0.0	0.0	0.0	0.0	76.7	9.91	43.29	46.88
G3g	389	546	7384	3.1	440.4	445.5	432.8	432.8	17.2	43.56	30.39	47.03
G3m	916	513	3872	8.8	410.2	412.9	402.2	402.2	73.3	43.80	58.56	47.03
G3o	943	234	1384	1.3	249.6	246.7	225.9	225.9	79.4	28.24	53.84	47.03
G3p	992	40	236	0.1	39.3	45.4	27.6	27.6	91.7	9.35	50.51	47.03
G3z	918	33	194	0.3	30.2	32.6	30.6	30.6	73.9	9.81	41.85	47.03
G2g	411	211	2858	1.2	118.7	123.7	118.1	118.1	21.1	28.19	24.65	47.47
G2m	1000	194	1466	3.3	110.3	115.4	110.1	110.1	100.0	29.07	64.54	47.47
G2o	992	117	702	0.7	79.3	74.4	71.0	71.0	92.8	19.40	56.09	47.47
G2p	996	11	66	0.0	5.5	7.6	7.0	7.0	94.4	4.63	49.54	47.47
G2z	940	17	100	0.2	8.5	8.3	8.0	8.0	78.3	6.76	42.55	47.47
B9g	591	5061	49270	17.2	6282.7	6359.2	6228.2	6228.2	32.8	95.79	64.28	47.92
B9m	50	4000	20000	36.6	5199.2	5375.0	5291.7	5291.8	1.7	89.40	45.53	47.92



Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk

B9o	284	1498	7934	6.7	2553.1	2661.8	2468.1	2468.1	8.9	65.32	37.11	47.92
B9p	733	655	3676	2.1	1046.9	1103.3	931.8	931.8	40.0	45.46	42.73	47.92
B9z	760	1061	5270	7.6	1083.5	984.2	936.4	936.4	43.9	55.97	49.93	47.92
B8g	571	4990	49068	17.1	6018.9	6139.0	6020.2	6020.3	31.7	94.91	63.29	48.66
B8m	50	4000	20000	36.6	5199.2	5375.0	5291.7	5291.8	1.1	89.95	45.53	48.66
B8o	296	1465	7810	6.6	2438.6	2535.2	2351.4	2351.4	11.7	63.70	37.69	48.66
B8p	777	599	3380	1.9	869.7	908.7	749.7	749.7	46.7	42.87	44.77	48.66
B8z	813	990	5068	7.4	819.7	763.9	728.5	728.5	51.7	52.41	52.04	48.66
G9g	778	4746	48140	20.1	6035.9	6223.7	6120.4	6120.5	47.2	94.91	71.06	48.98
G9m	50	4000	20000	45.7	5199.2	5375.0	5291.7	5291.8	2.8	90.79	46.78	48.98
G9o	285	1313	7214	6.8	2336.5	2818.2	2404.2	2404.2	9.4	63.47	36.46	48.98
G9p	764	411	2356	1.5	735.2	1172.6	795.1	795.1	44.4	39.72	42.08	48.98
G9z	779	746	4140	6.9	836.6	848.7	828.7	828.7	47.8	49.21	48.50	48.98
G5g	395	2899	34516	14.4	2933.2	3140.6	3073.5	3073.5	18.9	74.58	46.74	49.22
G5m	834	2809	16918	38.6	2820.8	3007.5	2947.3	2947.3	55.0	74.81	64.91	49.22
G5o	648	631	3698	3.5	886.8	1080.0	914.1	914.1	33.9	46.34	40.12	49.22
G5p	972	116	680	0.4	152.8	285.0	152.0	152.0	87.2	19.77	53.50	49.22
G5z	876	90	502	0.8	112.4	133.1	126.2	126.2	62.2	19.44	40.83	49.22
G8g	858	4283	45544	19.0	5680.4	5891.1	5792.0	5792.0	59.4	93.06	76.25	49.33
G8m	50	4000	20000	45.7	5199.2	5375.0	5291.7	5291.8	2.2	91.34	46.78	49.33
G8o	253	1244	6952	6.6	2195.5	2588.7	2239.8	2239.8	6.7	61.06	33.87	49.33
G8p	825	322	1856	1.2	543.6	893.3	581.1	581.1	53.3	34.40	43.87	49.33
G8z	849	283	1544	2.6	481.2	516.1	500.2	500.2	57.2	34.58	45.90	49.33
G6g	464	3288	37942	15.9	3734.2	3940.4	3854.4	3854.4	27.2	81.34	54.28	49.49
G6m	694	3152	17984	41.1	3547.3	3716.7	3641.7	3641.7	36.7	81.25	58.96	49.49
G6o	471	875	5064	4.8	1394.0	1674.8	1427.8	1427.8	27.8	52.87	40.32	49.49
G6p	925	176	1026	0.6	249.4	475.2	262.5	262.5	75.6	25.23	50.39	49.49
G6z	876	136	766	1.3	186.9	223.8	212.6	212.6	62.8	24.21	43.50	49.49
B7g	667	4886	48578	17.0	5678.9	5814.1	5707.1	5707.1	34.4	93.52	63.98	49.58
B7m	78	3967	19934	36.4	5028.6	5226.8	5146.7	5146.8	3.3	88.06	45.69	49.58
B7o	289	1368	7430	6.3	2220.1	2300.8	2129.8	2129.9	11.1	61.94	36.53	49.58
B7p	854	478	2720	1.6	593.7	607.4	465.2	465.2	58.9	38.52	48.70	49.58
B7z	842	919	4842	7.0	650.2	587.3	560.3	560.3	55.6	50.46	53.01	49.58
B6g	688	4761	47670	16.6	5139.1	5323.3	5230.6	5230.7	35.6	90.42	62.99	49.76
B6m	160	3855	19644	35.9	4552.6	4795.1	4728.2	4728.2	3.9	85.93	44.91	49.76
B6o	320	1287	7044	6.0	1922.4	2011.5	1852.4	1852.4	13.3	59.49	36.41	49.76
B6p	899	420	2406	1.4	428.6	432.0	299.3	299.3	67.8	35.05	51.41	49.76
B6z	847	906	4796	7.0	586.5	528.2	502.5	502.5	56.7	49.49	53.08	49.76
G7g	880	3867	42628	17.8	4839.9	5078.1	4985.6	4985.7	63.9	87.08	75.49	50.25

G7m	354	3681	19362	44.2	4546.8	4745.6	4665.6	4665.7	15.6	85.19	50.37	50.25
G7o	246	1181	6710	6.3	2062.7	2430.6	2101.8	2101.8	6.1	59.54	32.82	50.25
G7p	881	243	1408	0.9	371.5	695.6	403.5	403.5	64.4	30.23	47.34	50.25
G7z	875	186	1044	1.7	293.1	332.4	320.0	320.0	61.7	28.80	45.23	50.25
B5g	691	4581	46568	16.3	4524.6	4697.5	4608.6	4608.6	36.1	87.08	61.60	50.36
B5m	278	3694	19288	35.3	4007.6	4247.4	4181.1	4181.1	8.3	83.01	45.67	50.36
B5o	393	1207	6634	5.6	1633.6	1724.0	1576.5	1576.4	17.8	57.31	37.55	50.36
B5p	915	394	2280	1.3	351.4	363.6	241.9	241.9	72.8	33.43	53.10	50.36
B5z	861	887	4728	6.9	517.0	450.2	427.5	427.5	60.0	47.78	53.89	50.36
B4g	824	3516	37970	13.3	2937.1	2998.9	2937.5	2937.5	52.2	78.94	65.58	52.55
B4m	446	2650	15784	28.9	2502.6	2625.1	2585.6	2585.6	25.6	71.20	48.38	52.55
B4o	522	818	4780	4.0	1065.1	1149.1	1032.3	1032.3	29.4	49.77	39.61	52.55
B4p	974	119	704	0.4	132.8	188.1	90.1	90.1	87.8	19.44	53.61	52.55
B4z	882	866	4640	6.7	434.5	373.8	351.9	351.9	65.0	46.11	55.56	52.55
B3g	825	3150	34974	12.2	2053.0	2066.9	2017.7	2017.7	53.9	73.61	63.75	52.71
B3m	576	2294	14336	26.2	1661.5	1742.0	1713.7	1713.7	32.2	67.13	49.68	52.71
B3o	710	606	3550	3.0	604.4	668.5	587.2	587.2	37.8	43.29	40.53	52.71
B3p	985	102	604	0.3	86.5	128.5	64.6	64.6	90.6	16.85	53.70	52.71
B3z	896	856	4598	6.7	391.5	324.8	304.0	304.0	66.7	45.09	55.88	52.71
P9g	394	6086.3	1858	0.9	5678.3	5499.0	6146.3	5676.6	18.3	78.70	48.52	54.83
P9m	733	3086.3	886	2.6	3125.2	3839.4	3243.1	3738.6	40.6	65.28	52.92	54.83
P9o	800	98.462	662	0.7	5318.1	6890.7	4916.0	4718.1	50.0	44.72	47.36	54.83
P9p	947	1012.0	248	0.2	5305.1	6112.1	5005.1	4164.7	81.1	51.67	66.39	54.83
P9z	907	3000	142	0.3	2553.1	1659.7	2903.2	1938.0	70.6	47.36	58.96	54.83
B2g	795	2738	31508	11.0	1197.4	1205.4	1185.2	1185.2	49.4	66.20	57.82	55.20
B2m	698	1949	12786	23.4	997.9	1042.3	1028.1	1028.1	37.2	63.47	50.35	55.20
B2o	921	336	2006	1.7	193.8	248.9	214.1	214.1	75.0	31.81	53.40	55.20
B2p	992	81	484	0.3	48.6	74.8	46.7	46.7	92.2	14.72	53.47	55.20
B2z	945	789	4346	6.3	199.5	163.1	157.1	157.1	80.6	41.34	60.95	55.20
P8g	399	6077	1748	0.9	5624.8	5464.3	6091.8	5641.8	20.0	77.73	48.87	55.42
P8m	758	3077	804	2.4	3077.3	3808.7	3193.4	3707.0	43.3	64.17	53.75	55.42
P8o	828	87.916	604	0.6	5276.6	6879.4	4871.7	4696.9	54.4	43.33	48.89	55.42
P8p	960	1010.9	230	0.2	5290.6	6110.1	4990.3	4155.4	83.9	50.74	67.31	55.42
P8z	904	3000	140	0.3	2547.5	1655.6	2898.4	1934.8	69.4	47.08	58.26	55.42
P7g	387	6068.8	1544	0.8	5564.2	5426.5	6031.0	5605.6	16.1	76.99	46.55	55.44
P7m	788	3068.8	706	2.1	3024.8	3779.0	3139.8	3678.1	48.3	62.96	55.65	55.44
P7o	850	78.787	522	0.6	5236.1	6861.3	4829.4	4671.0	57.8	42.08	49.93	55.44
P7p	958	1009.9	224	0.2	5272.2	6102.0	4974.0	4139.5	83.3	49.95	66.64	55.44
P7z	905	3000	134	0.3	2539.3	1647.5	2891.3	1927.4	70.0	46.81	58.40	55.44

Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk

P6g	410	6059.3	1426	0.7	5503.8	5367.2	5971.6	5544.8	20.6	75.74	48.15	56.37
P6m	802	3059.3	668	2.0	2971.8	3736.8	3086.7	3633.4	50.6	61.90	56.23	56.37
P6o	862	68.418	480	0.5	5188.4	6845.2	4782.9	4645.1	60.6	40.60	50.58	56.37
P6p	966	1009.0	186	0.1	5255.8	6107.8	4962.6	4137.0	86.1	49.03	67.57	56.37
P6z	910	3000	118	0.2	2532.0	1630.5	2884.9	1911.4	72.2	46.39	59.31	56.37
B1g	537	1950	23888	8.3	0.0	0.0	0.0	0.0	31.1	51.76	41.44	57.54
B1m	1000	1302	9198	16.8	0.0	0.0	0.0	0.0	96.7	49.72	73.19	57.54
B1o	1000	146	876	0.7	0.0	0.0	0.0	0.0	97.2	18.80	58.01	57.54
B1p	1000	59	354	0.2	0.0	0.0	0.0	0.0	97.8	9.63	53.70	57.54
B1z	970	648	3770	5.5	0.0	0.0	0.0	0.0	86.7	36.02	61.34	57.54
P5g	421	6050.2	1228	0.6	5445.3	5288.2	5912.9	5466.1	22.2	74.40	48.31	58.09
P5m	842	3050.2	578	1.7	2922.4	3677.5	3035.1	3575.2	56.1	60.46	58.29	58.09
P5o	903	58.17	432	0.5	5143.5	6815.8	4739.3	4610.1	68.9	39.58	54.24	58.09
P5p	975	1007.9	182	0.1	5234.6	6106.0	4946.1	4130.8	88.9	48.38	68.63	58.09
P5z	925	3000	98	0.2	2522.9	1610.7	2877.9	1891.0	76.1	45.83	60.97	58.09
P4g	429	6041.6	954	0.5	5379.5	5225.4	5848.4	5402.8	25.0	72.78	48.89	58.78
P4m	880	3041.6	468	1.4	2866.0	3629.6	2978.1	3526.8	63.3	58.94	61.13	58.78
P4o	908	48.183	376	0.4	5092.3	6785.8	4690.5	4571.3	71.7	37.45	54.56	58.78
P4p	974	1006.5	164	0.1	5214.8	6105.3	4932.2	4122.9	88.3	47.64	67.99	58.78
P4z	939	3000	66	0.1	2513.5	1595.8	2870.3	1876.0	77.8	44.91	61.34	58.78
P3g	446	6034.1	842	0.4	5316.4	5186.4	5786.0	5363.3	26.1	71.76	48.94	59.63
P3m	896	3034.1	406	1.2	2812.7	3592.2	2923.5	3488.7	67.2	57.45	62.34	59.63
P3o	933	39.313	322	0.3	5041.3	6753.9	4641.7	4531.2	77.2	35.93	56.57	59.63
P3p	979	1005.1	142	0.1	5191.0	6091.0	4915.4	4101.6	89.4	46.71	68.08	59.63
P3z	944	3000	60	0.1	2503.7	1594.2	2862.5	1874.6	80.0	44.40	62.20	59.63
P2g	419	6026.2	706	0.4	5248.3	5142.6	5718.4	5320.7	21.7	70.23	45.95	60.30
P2m	920	3026.2	340	1.0	2754.3	3557.1	2863.8	3454.4	74.4	56.06	65.25	60.30
P2o	951	29.846	242	0.3	4986.8	6724.9	4588.8	4494.5	81.7	33.94	57.80	60.30
P2p	992	1003.5	106	0.1	5165.3	6082.3	4896.6	4084.5	93.3	45.60	69.47	60.30
P2z	953	3000	42	0.1	2494.0	1585.5	2854.6	1866.2	82.2	43.84	63.03	60.30
P1g	422	6013.3	360	0.2	5172.1	5048.1	5642.1	5226.8	22.8	65.79	44.28	61.36
P1m	965	3013.3	178	0.5	2687.4	3463.2	2794.8	3360.6	85.6	52.45	69.00	61.36
P1o	987	15.666	126	0.1	4920.5	6659.7	4526.3	4422.0	91.1	30.74	60.93	61.36
P1p	996	1002.2	50	0.0	5133.5	6071.4	4874.0	4067.7	95.0	44.31	69.65	61.36
P1z	956	3000	16	0.0	2484.7	1584.9	2847.4	1866.2	82.8	43.10	62.94	61.36